Smoothed Analysis of Online and Differentially Private Learning

Nika Haghtalab * Tim Roughgarden † Abhishek Shetty ‡

July 25, 2020

Abstract

Practical and pervasive needs for robustness and privacy in algorithms have inspired the design of online adversarial and differentially private learning algorithms. The primary quantity that characterizes learnability in these settings is the Littlestone dimension of the class of hypotheses [**??**]. This characterization is often interpreted as an impossibility result because classes such as linear thresholds and neural networks have infinite Littlestone dimension. In this paper, we apply the framework of smoothed analysis [**?**], in which adversarially chosen inputs are perturbed slightly by nature. We show that fundamentally stronger regret and error guarantees are possible with smoothed adversaries than with worst-case adversaries. In particular, we obtain regret and privacy error bounds that depend only on the VC dimension and the bracketing number of a hypothesis class, and on the magnitudes of the perturbations.

1 Introduction

Robustness to changes in the data and protecting the privacy of data are two of the main challenges faced by machine learning and have led to the design of *online* and *differentially private* learning algorithms. While offline PAC learnability is characterized by the finiteness of VC dimension, online and differentially private learnability are both characterized by the finiteness of the Littlestone dimension [**???**]. This latter characterization is often interpreted as an impossibility result for achieving robustness and privacy on worst-case instances, especially in classification where even simple hypothesis classes such as 1-dimensional thresholds have constant VC dimension but infinite Littlestone dimension.

Impossibility results for worst-case adversaries do not invalidate the original goals of robust and private learning with respect to practically relevant hypothesis classes; rather, they indicate that a new model is required to provide rigorous guidance on the design of online and differentially private learning algorithms. In this work, we go beyond worst-case analysis and *design online learning algorithms and differentially private learning algorithms as good as their offline and non-private PAC learning counterparts in a realistic semi-random model of data.*

Inspired by smoothed analysis [**?**], we introduce frameworks for online and differentially private learning in which adversarially chosen inputs are perturbed slightly by nature (reflecting, e.g., measurement errors or uncertainty). Equivalently, we consider an adversary restricted to choose an input distribution that is not overly concentrated, with the realized input then drawn from the adversary's chosen distribution. Our goal is to design algorithms with good expected regret and error bounds, where the expectation is over nature's perturbations (and any random coin flips of the algorithm). Our positive results show, in a precise sense, that the known lower bounds for worst-case online and differentially private learnability are fundamentally brittle.

Our Model. Let us first consider the standard online learning setup with an instance space $\mathcal{X}$ and a set $\mathcal{H}$ of binary hypotheses each mapping $\mathcal{X}$ to $\mathcal{Y} = \{+1, -1\}$. Online learning is played over T time steps, where

*Cornell University; Email: nika@cs.cornell.edu
†Columbia University; Email: tr@cs.columbia.edu
‡Cornell University; Email: shetty@cs.cornell.edu

at each step the learner picks a prediction function from a distribution and the *adaptive* adversary chooses a pair of $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$. The regret of an algorithm is the difference between the number of mistakes the algorithm makes and that of the best fixed hypothesis in $\mathcal{H}$. The basic goal in online learning is to obtain a regret of $o(T)$. In comparison, in differential privacy the data set $B = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is specified ahead of time. Our goal here is to design a randomized mechanism that with high probability finds a nearly optimal hypothesis in $\mathcal{H}$ on the set B, while ensuring that the computation is *differentially private*. That is, changing a single element of B does not significantly alter the probability with which our mechanism selects an outcome. Similar to agnostic PAC learning, this can be done by ensuring that the error of each hypothesis $h \in \mathcal{H}$ on B (referred to as a query) is calculated accurately and privately.

We extend these two models to accommodate smoothed adversaries. We say that a distribution $\mathcal{D}$ over instance-label pairs is *σ-smooth* if its density function over the instance domain is pointwise bounded by at most $1/\sigma$ times that of the uniform distribution. In the online learning setting this means that at step t, the adversary chooses an arbitrary σ-smooth distribution $\mathcal{D}_t$ from which $(x_t, y_t) \sim \mathcal{D}_t$ is drawn. In the differential privacy setting, we work with a database B for which the answers to the queries could have been produced by a σ-smooth distribution.

Why should smoothed analysis help in online learning? Consider the well-known lower bound for 1-dimensional thresholds over $\mathcal{X} = [0, 1]$, in which the learner may as well perform binary search and the adversary selects an instance within the uncertainty region of the learner that causes a mistake. While the learner's uncertainty region is halved each time step, the worst-case adversary can use ever-more precision to force the learner to make mistakes indefinitely. On the other hand, a σ-smoothed adversary effectively has bounded precision. That is, once the width of the uncertainty region drops below σ, a smoothed adversary can no longer guarantee that the chosen instance lands in this region. Similarly for differential privacy, there is a σ-smooth distribution that produces the same answers to the queries. Such a distribution has no more than α probability over an interval of width $\sigma\alpha$. So one can focus on computing the errors of the $1/(\sigma\alpha)$ hypotheses with discreized thresholds and learn a hypothesis of error at most α. Analogous observations have been made in prior works (**?**, **?**, **?**), although only for very specific settings (online learning of 1-dimensional thresholds, 1-dimensional piecewise constant functions, and parameterized greedy heuristics for the maximum weight independent set problem, respectively). Our work is the first to demonstrate the breadth of the settings in which fundamentally stronger learnability guarantees are possible for smoothed adversaries than for worst-case adversaries.

Our Results and Contributions.

- Our main result concerns online learning with *adaptive* σ-smooth adversaries where $\mathcal{D}_t$ can depend on the history of the play, including the earlier realizations of $x_\tau \sim \mathcal{D}_\tau$ for $\tau < t$. That is, x_t and $x_{t'}$ can be highly correlated. We show that regret against these powerful adversaries is bounded by $\tilde{O}(\sqrt{T \ln(\mathcal{N})})$, where $\mathcal{N}$ is the *bracketing number* of $\mathcal{H}$ with respect to the uniform distribution.[1] Bracketing number is the size of an ϵ-cover of $\mathcal{H}$ with the additional property that hypotheses in the cover are *pointwise* approximations of those in $\mathcal{H}$. We show that for many hypothesis classes, the bracketing number is nicely bounded as a function of the VC dimension. This leads to the regret bound of $\tilde{O}(\sqrt{T \ \mathrm{VCDim}(\mathcal{H}) \ln(1/\sigma)})$ for commonly used hypothesis classes in machine learning, such as halfspaces, polynomial threshold functions, and polytopes. In comparison, these hypothesis classes have infinite Littlestone dimension and thus cannot be learned with regret $o(T)$ in the worst case [**?**].

  From a technical perspective, we introduce a novel approach for bounding time-correlated non-independent stochastic processes over infinite hypothesis classes using the notion of bracketing number. Furthermore, we introduce systematic approaches, such as high-dimensional linear embeddings and k-fold operations, for analyzing the bracketing number of complex hypothesis classes. We believe these techniques are of independent interest.

- For differentially private learning, we obtain an error bound of $\tilde{O}\big(\ln^{\frac{3}{8}}(1/\sigma)\sqrt{\mathrm{VCDim}(\mathcal{H})/n}\big)$; the key point is that this bound is independent of the size $|\mathcal{X}|$ of the domain and the size $|\mathcal{H}|$ of the hypothesis class. We obtain these bounds by modifying two commonly used mechanisms in differential

[1] Along the way, we also demonstrate a stronger regret bound for the simpler case of non-adaptive adversaries, for which each distribution $\mathcal{D}_t$ is independent of the realized inputs in previous time steps.

privacy, the Multiplicative Weight Exponential Mechanism of **?** and the SmallDB algorithm of **?**. With worst-case adversaries, these algorithms achieve only error bounds of $\tilde{O}(\ln^{\frac{1}{4}}(|\mathcal{X}|)\sqrt{\ln(|\mathcal{H}|)/n})$ and $\tilde{O}(\sqrt[3]{\text{VCDim}(\mathcal{H})\ln(|\mathcal{X}|)/n})$, respectively. Our results also improve over those in **?** which concern a similar notion of smoothness and achieve an error bound of $\tilde{O}(\ln^{\frac{1}{2}}(1/\sigma)\sqrt{\ln(|\mathcal{H}|)/n})$.

Other Related Works. At a higher level, our work is related to several works on the intersection of machine learning and beyond the worst-case analysis of algorithms (e.g., [**???**]) that are covered in more detail in Appendix A.

2 Preliminaries

Online Learning. We consider a measurable instance space $\mathcal{X}$ and the label set $\mathcal{Y} = \{+1, -1\}$. Let $\mathcal{H}$ be a hypothesis class on $\mathcal{X}$ with its VC dimension denoted by $\text{VCDim}(\mathcal{H})$. Let $\mathcal{U}$ be the uniform distribution over $\mathcal{X}$ with density function $u(\cdot)$. For a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, let $p(\cdot)$ be the *probability density function* of its marginal over $\mathcal{X}$. We say that $\mathcal{D}$ is σ-*smooth* if for all $x \in \mathcal{X}$, $p(x) \leq u(x)\sigma^{-1}$. For a labeled pair $s = (x, y)$ and a hypothesis $h \in \mathcal{H}$, $\text{err}_s(h) = 1(h(x) \neq y)$ indicates whether h makes a mistake on s.

We consider the setting of *online adversarial and (full-information) learning*. In this setting, a learner and an adversary play a repeated game over T time steps. In every time step $t \in [T]$ the learner picks a hypothesis h_t and adversary picks a σ-smoothed distribution $\mathcal{D}_t$ from which a labeled pair $s_t = (x_t, y_t)$ such that $s_t \sim \mathcal{D}_t$ is generated. The learner then incurs penalty of $\text{err}_{s_t}(h_t)$. We consider two types of adversaries. First (and the subject of our main results) is called an *adaptive* σ-smooth adversary. This adversary at every time step $t \in [T]$ chooses $\mathcal{D}_t$ based on the actions of the learner $h_1, \ldots, h_{t-1}$ and, importantly, the realizations of the previous instances $s_1, \ldots, s_{t-1}$. We denote this adaptive random process by $\mathbf{s} \sim \boldsymbol{\mathscr{D}}$. A second and less powerful type of adversary is called a *non-adaptive* σ-smooth adversary. Such an adversary first chooses an unknown sequence of distributions $\boldsymbol{\mathcal{D}} = (\mathcal{D}_1, \ldots, \mathcal{D}_T)$ such that $\mathcal{D}_t$ is a σ-smooth distribution for all $t \in [T]$. Importantly, $\mathcal{D}_t$ does not depend on realizations of adversary's earlier actions $s_1, \ldots, s_{t-1}$ or the learner's actions $h_1, \ldots, h_{t-1}$. We denote this non-adaptive random process by $\mathbf{s} \sim \boldsymbol{\mathcal{D}}$. With a slight abuse of notation, we denote by $\mathbf{x} \sim \boldsymbol{\mathscr{D}}$ and $\mathbf{x} \sim \boldsymbol{\mathcal{D}}$ the sequence of (unlabeled) instances in $\mathbf{s} \sim \boldsymbol{\mathscr{D}}$ and $\mathbf{s} \sim \boldsymbol{\mathcal{D}}$.

Our goal is to design an online algorithm $\mathcal{A}$ such that expected regret against an adaptive adversary,

$$\mathbb{E}[\textsc{Regret}(\mathcal{A}, \boldsymbol{\mathscr{D}})] := \mathop{\mathbb{E}}_{\mathbf{s}\sim\boldsymbol{\mathscr{D}}}\left[\sum_{t=1}^{T} \text{err}_{s_t}(h_t) - \min_{h\in\mathcal{H}} \sum_{t=1}^{T} \text{err}_{s_t}(h)\right]$$

is sublinear in T. We also consider the regret of an algorithm against a non-adaptive adversary defined similarly as above and denoted by $\mathbb{E}[\textsc{Regret}(\mathcal{A}, \boldsymbol{\mathcal{D}})]$.

Differential Privacy. We also consider differential privacy. In this setting, a data set S is a multiset of elements from domain $\mathcal{X}$. Two data sets S and S' are said to be *adjacent* if they differ in at most one element. A randomized algorithm $\mathcal{M}$ that takes as input a data set is (ϵ, δ)-differentially private if for all $\mathcal{R} \subseteq \text{Range}(\mathcal{M})$ and for all adjacent data sets S and S', $\Pr[\mathcal{M}(S) \in \mathcal{R}] \leq \exp(\epsilon)\Pr[\mathcal{M}(S') \in \mathcal{R}] + \delta$. If $\delta = 0$, the algorithm is said to be purely ϵ-differentially private.

For differentially private learning, one considers a fixed class of *queries* $\mathcal{Q}$. The learner's goal is to evaluate these queries on a given data set S. For ease of notation, we work with the empirical distribution $\mathcal{D}_S$ corresponding to a data set S. Then the learner's goal is to approximately compute $q(\mathcal{D}_S) = \mathbb{E}_{x\sim\mathcal{D}_S}[q(x)]$ while preserving privacy[2]. We consider two common paradigms of differential privacy. First, called *query answering*, involves designing a mechanism that outputs values v_q for all $q \in \mathcal{Q}$ such that with probability $1-\beta$ for every $q \in Q$, $|q(\mathcal{D}_S) - v_q| \leq \alpha$. The second paradigm, called *data release*, involves designing a mechanism that outputs a synthetic distribution $\overline{\mathcal{D}}$, such that with probability $1-\beta$ for all $q \in \mathcal{Q}$, $|q(\overline{\mathcal{D}}) - q(\mathcal{D}_S)| \leq \alpha$. That is, the user can use $\overline{\mathcal{D}}$ to compute the value of any $q(\mathcal{D}_S)$ approximately.

Analogous to the definition of smoothness in online learning, we say that a distribution $\mathcal{D}$ with density function $p(\cdot)$ is σ-*smooth* if $p(x) \leq \sigma^{-1}u(x)$ for all $x \in \mathcal{X}$. We also work with a weaker notion of smoothness

[2]In differentially private learning, queries are the error function of hypotheses and take as input a pair (x, y).

of data sets. A data set S is said to be (σ, χ)-*smooth* with respect to a query set $\mathcal{Q}$ if there is a σ-smooth distribution $\mathcal{D}$ such that for all $q \in \mathcal{Q}$, we have $|q(\mathcal{D}) - q(\mathcal{D}_S)| \leq \chi$. The definition of (σ, χ)-smoothness, which is also referred to as *pseudo-smoothness* by **?**, captures data sets that though might be concentrated on some elements, the query class is not capable of noticing their lack of smoothness.

Additional Definitions. Let $\mathcal{H}$ be a hypothesis class and let $\mathcal{D}$ be a distribution. $\mathcal{H}'$ is an ϵ-cover for $\mathcal{H}$ under $\mathcal{D}$ if for all $h \in \mathcal{H}$, there is a $h' \in \mathcal{H}'$ such that $\Pr_{x\sim\mathcal{D}}[h(x) \neq h'(x)] \leq \epsilon$. For any $\mathcal{H}$ and $\mathcal{D}$, there an ϵ-cover $\mathcal{H}' \subseteq \mathcal{H}$ under $\mathcal{D}$ such that $|\mathcal{H}'| \leq (41/\epsilon)^{\text{VCDim}(\mathcal{H})}$ (**?**).

We define a partial order $\preceq$ over functions such that $f_1 \preceq f_2$ if and only if for all $x \in \mathcal{X}$, we have $f_1(x) \leq f_2(x)$. For a pair of functions f_1, f_2 such that $f_1 \preceq f_2$, a *bracket* $[f_1, f_2]$ is defined by $[f_1, f_2] = \{f : \mathcal{X} \to \{-1, 1\} : f_1 \preceq f \preceq f_2\}$. Given a measure μ over $\mathcal{X}$, a bracket $[f_1, f_2]$ is called an ϵ-bracket if $\Pr_{x\sim\mu}[f_1(x) \neq f_2(x)] \leq \epsilon$.

Definition 2.1 (Bracketing Number)**.** *Consider an instance space $\mathcal{X}$, measure μ over this space, and hypothesis class $\mathcal{F}$. A set $\mathcal{B}$ of brackets is called an* ϵ-bracketing of $\mathcal{F}$ with respect to measure μ *if all brackets in $\mathcal{B}$ are ϵ-brackets with respect to μ and for every $f \in \mathcal{F}$ there is $[f_1, f_2] \in \mathcal{B}$ such that $f \in [f_1, f_2]$. The* ϵ-bracketing number *of $\mathcal{F}$ with respect to measure μ, denoted by $\mathcal{N}_{[]}(\mathcal{F}, \mu, \epsilon)$, is the size of the smallest ϵ-bracketing for $\mathcal{F}$ with respect to μ.*

3 Regret Bounds for Smoothed Adaptive and Non-Adaptive Adversaries

In this section, we obtain regret bounds against smoothed adversaries. For finite hypothesis classes $\mathcal{H}$, existing no-regret algorithms such as Hedge [**?**] and Follow-the-Perturbed-Leader [**?**] achieve a regret bound of $O(\sqrt{T \ln(\mathcal{H})})$. For a possibly infinite hypothesis class our approach uses a finite set $\mathcal{H}'$ as a *proxy* for $\mathcal{H}$ and only focuses on competing with hypotheses in $\mathcal{H}'$ by running a standard no-regret algorithm on $\mathcal{H}'$. Indeed, in absence of smoothness of $\mathscr{D}$, $\mathcal{H}'$ has to be a good proxy with respect to every distribution or know the adversarial sequence ahead of time, neither of which are possible in the online setting. But when distributions are smooth, $\mathcal{H}'$ that is a good proxy for the uniform distribution can also be a good proxy for all other smooth distributions. We will see that how well a set $\mathcal{H}'$ approximates $\mathcal{H}$ depends on adaptivity (versus non-adpativity) of the adversary. Our main technical result in Section 3.1 shows that for adaptive adversaries this approximation depends on the size of the $\frac{\sigma}{4\sqrt{T}}$-bracketing cover of $\mathcal{H}$. This results in an algorithm whose regret is sublinear in T and logarithmic in that bracketing number for adaptive adversaries (Theorem 3.3). In comparison, for simpler non-adaptive adversaries this approximation depends on the size of the more traditional ϵ-covers of $\mathcal{H}$, which do not require pointwise approximation of $\mathcal{H}$. This leads to an algorithm against non-adaptive adversaries with an improved regret bound of $\tilde{O}(\sqrt{T \cdot \text{VCDim}(\mathcal{H})})$ (Theorem 3.3).

In Section 3.2, we demonstrate that the bracketing numbers of commonly used hypothesis classes in machine learning are small functions of their VC dimension. We also provide systematic approaches for bounding the bracketing number of complex hypothesis classes in terms of the bracketing number of their simpler building blocks. This shows that for many commonly used hypothesis classes — such as halfspaces, polynomial threshold functions, and polytopes — we can achieve a regret of $\tilde{O}(\sqrt{T \cdot \text{VCDim}(\mathcal{H})})$ even against an adaptive adversary.

3.1 Regret Analysis and the Connection to Bracketing Number

In more detail, consider an algorithm $\mathcal{A}$ that uses Hedge on a finite set $\mathcal{H}'$ instead of $\mathcal{H}$. Then,

$$\mathbb{E}[\textsc{Regret}(\mathcal{A}, \mathscr{D})] \leq O\left(\sqrt{T \ln(|\mathcal{H}'|)}\right) + \mathop{\mathbb{E}}_{\mathscr{D}}\left[\max_{h\in\mathcal{H}} \min_{h'\in\mathcal{H}'} \sum_{t=1}^{T} 1\left(h(x_t) \neq h'(x_t)\right)\right], \tag{1}$$

where the first term is the regret against the best $h' \in \mathcal{H}'$ and the second term captures how well $\mathcal{H}'$ approximates $\mathcal{H}$. A natural choice of $\mathcal{H}'$ is an ϵ-cover of $\mathcal{H}$ with respect to the uniform distribution, for a small ϵ that will be defined later. This bounds the first term using the fact that there is an ϵ-cover $\mathcal{H}' \subseteq \mathcal{H}$ of

size $|\mathcal{H}'| \leq (41/\epsilon)^{\text{VCDim}(\mathcal{H})}$. To bound the second term, we need to understand whether there is a hypothesis $h \in \mathcal{H}$ whose value over *an adaptive sequence of σ-smooth distributions* can be drastically different from the value of its closest (under uniform distribution) proxy $h' \in \mathcal{H}'$. Considering the symmetric difference functions $f_{h,h'} = h \Delta h'$ for functions $h \in \mathcal{H}$ and their corresponding proxies $h' \in \mathcal{H}'$, we need to bound (in expectation) the maximum value an $f_{h,h'}$ can attain over an adaptive sequence of σ-smooth distributions.

Non-Adaptive Adversaries. To develop more insight, let us first consider the case of *non-adaptive* adversaries. In the case of non-adaptive adversaries, $x_t \sim \mathcal{D}_t$ are *independent* of each other, while they are not identically distributed. This independence is the key property that allows us to use the VC dimension of the set of functions $\{f_{h,h'} \mid \forall h \in \mathcal{H}$ and the corresponding proxy $h' \in \mathcal{H}'\}$ to establish a uniform convergence property where with high probability every function $f_{h,h'}$ has a value that is close to its expectation — the fact that x_ts are not identically distributed can be easily handled because the double sampling and symmetrization trick in VC theory can still be applied as before. Furthermore, σ-smoothness of the distributions implies that $\mathbb{E}_{\boldsymbol{\mathcal{D}}}[\sum f_{h,h'}(x_t)] \leq \sigma^{-1}\, \mathbb{E}_{\mathcal{U}}[\sum f_{h,h'}(x_t)] \leq \epsilon/\sigma$. This leads to the following theorem for non-adaptive adversaries.

Theorem 3.1 (Non-Adaptive Adversary [**?**])**.** *Let $\mathcal{H}$ be a hypothesis class of VC dimension d. There is an algorithm such that for any $\boldsymbol{\mathcal{D}}$ that is an non-adaptive sequence of σ-smooth distributions has regret* $\mathbb{E}[\textsc{Regret}(\mathcal{A}, \boldsymbol{\mathcal{D}})] \in O\left(\sqrt{dT \ln\left(\frac{T}{\sigma}\right)}\right)$.

Adaptive Adversaries. Moving back to the case of adaptive adversaries, we unfortunately lose this uniform convergence property (see Appendix B for an example). This is due to the fact that now the choice of $\mathcal{D}_t$ can depend on the earlier realization of instances $x_1, \ldots, x_{t-1}$. To see why independence is essential, note that the ubiquitous double sampling and symmetrization techniques used in VC theory require that taking two sets of samples $\mathbf{x}$ and $\mathbf{x}'$ from the process that is generating data, we can swap x_i and x'_i independently of whether x_j and x'_j are swapped for $j \neq i$. When the choice of $\mathcal{D}_t$ depends on $x_1, \ldots, x_{t-1}$ then swapping x_τ with x'_τ affects whether x_t and x'_t could even be generated from $\mathcal{D}_t$ for $t > \tau$. In other words, symmetrizing the first t variables generates 2^t possible choices for x^{t+1} that exponentially increases the set of samples over which a VC class has to be projected, therefore losing the typical $\sqrt{T \cdot \text{VCDim}(\mathcal{H})}$ regret bound and instead obtaining the trivial regret of $O(T)$. Nevertheless, we show that the earlier ideas for bounding the second term of Equation 1 are still relevant as long as we can side step the need for independence.

Note that σ-smoothness of the distributions still implies that for a fixed function $f_{h,h'}$ even though $\mathcal{D}_t$ is dependent on the realizations $x_1, \ldots, x_{t-1}$, we still have $\Pr_{x_t \sim \mathcal{D}_t}[f_{h,h'}(x_t)] \leq \epsilon/\sigma$. Indeed, the value of any function f for which $\mathbb{E}_{\mathcal{U}}[f(x)] \leq \epsilon$ can be bounded by the convergence property of an appropriately chosen Bernoulli variable. As we demonstrate in the following lemma, this allows us to bound the expected maximum value of a $f_{h,h'}$ *chosen from a finite set of symmetric differences.* For a proof of this lemma refer to Appendix C.2.

Lemma 3.2. *Let $\mathcal{F} : \mathcal{X} \to \{0, 1\}$ be any finite class of functions such that $\mathbb{E}_{\mathcal{U}}[f(x)] \leq \epsilon$ for all $f \in \mathcal{F}$, i.e., every function has measure ϵ over the uniform distribution. Let $\boldsymbol{\mathscr{D}}$ be any adaptive sequence of T, σ-smooth distributions for some $\sigma \geq \epsilon$ such that $T\frac{\epsilon}{\sigma} \geq \sqrt{\ln(|\mathcal{F}|)}$. We have that*

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \boldsymbol{\mathscr{D}}}\left[\max_{f \in \mathcal{F}} \sum_{t=1}^{T} f(x_t)\right] \leq O\left(T\frac{\epsilon}{\sigma}\sqrt{\ln(|\mathcal{F}|)}\right).$$

The set of symmetric differences $\mathcal{G} = \{f_{h,h'} \mid \forall h \in \mathcal{H}$ and the corresponding proxy $h' \in \mathcal{H}'\}$ we work with is of course infinitely large. Therefore, to apply Lemma 3.2 we have to approximate $\mathcal{G}$ with a finite set $\mathcal{F}$ such that

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \boldsymbol{\mathscr{D}}}\left[\max_{f_{h,h'} \in \mathcal{G}} \sum_{t=1}^{T} f_{h,h'}(x_t)\right] \lesssim \mathop{\mathbb{E}}_{\mathbf{x} \sim \boldsymbol{\mathscr{D}}}\left[\max_{f \in \mathcal{F}} \sum_{t=1}^{T} f(x_t)\right]. \tag{2}$$

What should this set $\mathcal{F}$ be? Note that choosing $\mathcal{F}$ that is an ϵ-cover of $\mathcal{G}$ under the uniform distribution is an ineffective attempt plagued by the the same lack of independence that we are trying to side step. In fact, while all functions $f_{h,h'}$ are ϵ close to the constant 0 functions with respect to the uniform distribution,

they are activated on different parts of the domain. So it is not clear that an adaptive adversary, who can see the earlier realizations of instances, cannot ensure that one of these regions will receive a large number realized instances. But a second look at Equation 2 suffices to see that this is precisely what we can obtain if $\mathcal{F}$ were to be the set of (upper) functions in an ϵ-bracketing of $\mathcal{G}$. That is, for every function $f_{h,h'} \in \mathcal{G}$ there is a function $f \in \mathcal{F}$ such that $f_{h,h'} \preceq f$. This proves Equation 2 with an exact inequality using the fact that pointwise approximation $f_{h,h'} \preceq f$ implies that the value of $f_{h,h'}$ is bounded by that of f for any set of instances $x_1, \ldots, x_T$ that could be generated by $\mathscr{D}$. Furthermore, functions in $\mathcal{G}$ are within ϵ of the constant 0 function over the uniform distribution, so $\mathcal{F}$ meets the criteria of Lemma 3.2 with the property that for all $f \in \mathcal{F}$, $\mathbb{E}_{\mathcal{U}}[f(x)] \leq \epsilon$. It remains to bound the size of class $|\mathcal{F}|$ in terms of the bracketing number of $\mathcal{H}$. This can be done by showing that the bracketing number of class $\mathcal{G}$, that is the class of all symmetric differences in $\mathcal{H}$, is approximately bounded by the same bracketing number of $\mathcal{H}$ (See Theorem 3.7 for more details). Putting these all together we get the following regret bound against smoothed adaptive adversaries.

Theorem 3.3 (Adaptive Adversary)**.** *Let $\mathcal{H}$ be a hypothesis class over domain $\mathcal{X}$, whose ϵ-bracketing number with respect to the uniform distribution over $\mathcal{X}$ is denoted by $\mathcal{N}_{[]}(\mathcal{H}, \mathcal{U}, \epsilon)$. There is an algorithm such that for any $\mathscr{D}$ that is an adaptive sequence of σ-smooth distributions has regret*

$$\mathbb{E}[\textsc{Regret}(\mathcal{A}, \mathscr{D})] \in O\left(\sqrt{T \ln\left(\mathcal{N}_{[]}\left(\mathcal{H}, \mathcal{U}, \frac{\sigma}{4\sqrt{T}}\right)\right)}\right).$$

3.2 Hypothesis Classes with Small Bracketing Numbers.

In this section, we analyze bracketing numbers of some commonly used hypothesis classes in machine learning. We start by reviewing the bracketing number of halfspaces and provide two systematic approaches for extending this bound to other commonly used hypothesis classes. Our first approach bounds the bracketing number of any class using the dimension of the space needed to embed it as halfspaces. Our second approach shows that k-fold operations on any hypothesis class, such as taking the class of intersections or unions of all k hypotheses in a class, only mildly increase the bracketing number. Combining these two techniques allows us to bound the bracketing number of commonly used classifiers such as halfspaces, polytopes, polynomial threshold functions, etc.

The connection between bracketing number and VC theory has been explored in recent works. **??** showed that finite VC dimension class also have finite ϵ-bracketing number but **?** (see **?** for a modern presentation) showed the dependence on $1/\epsilon$ can be arbitrarily bad. Since Theorem 3.3 depends on the growth rate of bracketing numbers, we work with classes for which we can obtain ϵ-bracketing numbers with reasonable growth rate, those that are close to the size of standard ϵ-covers.

Theorem 3.4 (**?**)**.** *Let $\mathcal{H}$ be the class of halfspaces over $\mathbb{R}^d$. For any $\epsilon > 0$ and any measure μ over $\mathbb{R}^d$, $\mathcal{N}_{[]}(\mathcal{H}, \mu, \epsilon) \leq \left(\frac{d}{\epsilon}\right)^{O(d)}$.*

Our first technique uses this property of halfspaces to bound the bracketing number of any hypothesis class as a function of the dimension of the spaces needed to embed this class as halfpsaces.

Definition 3.5 (Embeddable Classes)**.** *Let $\mathcal{G}$ be a hypothesis class on $\mathcal{X}$. We say that $\mathcal{G}$ is embeddable as halfspaces in m dimensions if there exists a map $\psi : \mathcal{X} \to \mathbb{R}^m$ such that for any $g \in \mathcal{G}$, there is a linear threshold function h such $g = h \circ \psi$.*

Theorem 3.6 (Bracketing Number of Embeddable Classes)**.** *Let $\mathcal{G}$ be a hypothesis class embeddable as halfspaces in m dimensions. Then, for any measure ν, $\mathcal{N}_{[]}(\mathcal{G}, \nu, \epsilon) \leq \left(\frac{m}{\epsilon}\right)^{O(m)}$.*

Our second technique shows that combining k classes, by respectively taking intersections or unions of any k functions from them, only mildly increases their bracketing number.

Theorem 3.7 (Bracketing Number of k-fold Operations)**.** *Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be k hypothesis classes. Let $\mathcal{F}_1 \cdot \mathcal{F}_2 \cdots \mathcal{F}_k$ and $\mathcal{F}_1 + \mathcal{F}_2 + \cdots + \mathcal{F}_k$ be the class of all hypotheses that are intersections and unions of k functions $f_i \in \mathcal{F}_i$, respectively. Then,*

$$\mathcal{N}_{[]}(\mathcal{F}_1 \cdot \mathcal{F}_2 \cdots \mathcal{F}_k, \mu, k\epsilon) \leq \prod_{i \in [k]} \mathcal{N}_{[]}(\mathcal{F}_i, \mu, \epsilon)$$

and

$$\mathcal{N}_{[]}(\mathcal{F}_1 + \mathcal{F}_2 + \cdots + \mathcal{F}_k, \mu, k\epsilon) \leq \prod_{i\in[k]} \mathcal{N}_{[]}(\mathcal{F}_i, \mu, \epsilon)\,.$$

For any hypothesis class $\mathcal{F}$ *and* $\mathcal{G} = \{f\Delta f' \mid \text{ for all} f, f' \in \mathcal{F}\}$, $\mathcal{N}_{[]}(\mathcal{G}, \mu, 4\epsilon) \leq \left(\mathcal{N}_{[]}(\mathcal{F}, \mu, \epsilon)\right)^4$.

We now use our techniques for bounding the bracketing number of complex classes by the bracketing number of their simpler building blocks to show that online learning with an adaptive adversary on a class of halfspaces, polytopes, and polynomial threshold functions has $\tilde{O}(\sqrt{T\ \mathrm{VCDim}(\mathcal{H})})$ regret.

Corollary 3.8. *Consider instance space* $\mathcal{X} = \mathbb{R}^n$ *and let* μ *be an arbitrary measure on* $\mathcal{X}$. *Let* $\mathcal{P}^{n,d}$ *be the class of* d*-degree* polynomial thresholds *and* $\mathcal{Q}^{n,k}$ *be the class* k*-polytopes in* $\mathbb{R}^n$. *Then,*

$$\mathcal{N}_{[]}\left(\mathcal{P}^{n,d}, \mu, \epsilon\right) \leq \exp\left(c_1 n^d \ln\left(n^d/\epsilon\right)\right) \text{ and } \mathcal{N}_{[]}\left(\mathcal{Q}^{n,k}, \mu, \epsilon\right) \leq \exp\left(c_2 nk \ln\left(\frac{nk}{\epsilon}\right)\right),$$

for some constants c_1 *and* c_2. *Furthermore, there is an online algorithm whose regret against an adaptive* σ*-smoothed adversary on the class* $\mathcal{P}^{n,d}$ *and* $\mathcal{Q}^{n,k}$ *is respectively* $\tilde{O}(\sqrt{T \cdot \mathrm{VCDim}(\mathcal{P}^{n,d}) \ln(1/\sigma)})$ *and* $\tilde{O}(\sqrt{T \cdot \mathrm{VCDim}(\mathcal{Q}^{n,k}) \ln(1/\sigma)})$.

4 Differential Privacy

In this section, we consider smoothed analysis of differentially private learning in *query answering* and *data release* paradigms. We primarily focus on $(\sigma, 0)$-smooth distributions and defer the general case of (σ, χ)-smooth distributions to Appendix G. For finite query classes $\mathcal{Q}$ and small domains, existing differentially private mechanisms achieve an error bound that depends on $\ln(|\mathcal{Q}|)$ and $\ln(|\mathcal{X}|)$. We leverage smoothness of data sets to improve these dependencies to $\mathrm{VCDim}(\mathcal{Q})$ and $\ln(1/\sigma)$.

An Existing Algorithm. **?** introduced a practical algorithm for data release, called Multiplicative Weights Exponential Mechanism (MWEM). This algorithm works for a finite query class $\mathcal{Q}$ over a finite domain $\mathcal{X}$. Given an data set B and its corresponding empirical distribution $\mathcal{D}_B$, MWEM iteratively builds distributions $\mathcal{D}_t$ for $t \in [T]$, starting from $\mathcal{D}_1 = \mathcal{U}$ that is the uniform distribution over $\mathcal{X}$. At stage t, the algorithm picks a $q_t \in \mathcal{Q}$ that approximately maximizes the error $|q_t(\mathcal{D}_{t-1}) - q_t(\mathcal{D}_B)|$ using a differentially private mechanism (Exponential mechanism). Then data set $\mathcal{D}_{t-1}$ is updated using the multiplicative weights update rule $\mathcal{D}_t(x) \propto \mathcal{D}_{t-1}(x) \exp\left(q_t(x)(m_t - q_t(\mathcal{D}_{t-1}))/2\right)$ where m_t is a differentially private estimate (via Laplace mechanism) for the value $q_t(\mathcal{D}_B)$. The output of the mechanism is a data set $\overline{\mathcal{D}} = \frac{1}{T}\sum_{t\in[T]} \mathcal{D}_t$. The formal guarantees of the algorithm are as follows.

Theorem 4.1 (**?**). *For any data set* B *of size* n, *a finite query class* $\mathcal{Q}$, $T \in \mathbb{N}$ *and* $\epsilon > 0$, *MWEM is* ϵ*-differentially private and with probability at least* $1 - 2T/|\mathcal{Q}|$ *produces a distribution* $\overline{\mathcal{D}}$ *over* $\mathcal{X}$ *such that* $\max_{q\in\mathcal{Q}} \left\{\left|q\left(\overline{\mathcal{D}}\right) - q\left(\mathcal{D}_B\right)\right|\right\} \leq 2\sqrt{\frac{\log|\mathcal{X}|}{T}} + \frac{10T\log|\mathcal{Q}|}{\epsilon n}$.

The analysis of MWEM keeps track of the KL divergence $\mathrm{D}_{\mathrm{KL}}\left(\mathcal{D}_B \| \mathcal{D}_t\right)$ and shows that at time t this value decreases by approximately the error of query q_t. At a high level, $\mathrm{D}_{\mathrm{KL}}\left(\mathcal{D}_B \| \mathcal{D}_1\right) \leq \ln(|\mathcal{X}|)$. Moreover, KL divergence of any two distributions is non-negative. Therefore, error of any query $q \in \mathcal{Q}$ after T steps follows the above bound.

Query Answering. To design a private query answering algorithm for a query class $\mathcal{Q}$ without direct dependence on $\ln(|\mathcal{Q}|)$ and $\ln(|\mathcal{X}|)$ we leverage smoothness of distributions. Our algorithm called the Smooth Multiplicative Weight Exponential Mechanism (Smooth MWEM), given an infinite set of queries $\mathcal{Q}$, considers a γ-cover $\mathcal{Q}'$ under the uniform distribution. Then, it runs the MWEM algorithm with $\mathcal{Q}'$ as the query set and constructs an empirical distribution $\overline{\mathcal{D}}$. Finally, upon being requested an answer to a query $q \in \mathcal{Q}$, it responds with $q'(\overline{\mathcal{D}})$, where $q' \in \mathcal{Q}'$ is the closest query to q under the uniform distribution. This algorithm is presented in Appendix E. Note that $\mathcal{Q}'$ does not depend on the data set B. This is the key property that enables us to work with a finite γ-cover of $\mathcal{Q}$ and extend the privacy guarantees of MWEM to infinite query

classes. In comparison, constructing a γ-cover of $\mathcal{Q}$ with respect to the empirical distribution $\mathcal{D}_B$ uses private information.

Let us now analyze the error of our algorithm and outline the reasons it does not directly depend on $\ln(|\mathcal{Q}|)$ and $\ln(|\mathcal{X}|)$. Recall that from the $(\sigma, 0)$-smoothness, there is a distribution $\overline{\mathcal{D}_B}$ that is σ-smooth and $q(\mathcal{D}_B) = q(\overline{\mathcal{D}_B})$ for all $q \in \mathcal{Q}$. Furthermore, $\mathcal{Q}'$ can be taken to be a subset of $\mathcal{Q}$ and thus B is $(\sigma, 0)$-smooth with respect to $\mathcal{Q}'$. The approximation of $\mathcal{Q}$ by a γ-cover introduces error in addition to the error of Theorem 4.1. This error is given by $|q(\mathcal{D}_B) - q'(\mathcal{D}_B)| \leq 2 \cdot \Pr_{\mathcal{U}}[q'(x) \neq q(x)]\,\sigma^{-1} \leq 2\gamma/\sigma$. Note that $|\mathcal{Q}'| \leq (41/\gamma)^{\mathrm{VCDim}(\mathcal{Q})}$, therefore, this removes the error dependence on the size of the query set $\mathcal{Q}$ while adding a small error of $2\gamma/\sigma$. Furthermore, Theorem 4.1 dependence on $\ln(|\mathcal{X}|)$ is due to the fact that for a worst-case (non-smooth) data set B, $\mathrm{D}_{\mathrm{KL}}(\mathcal{D}_B \| \mathcal{U})$ can be as high as $\ln(|\mathcal{X}|)$. For a $(\sigma, 0)$-smooth data set, however, $\mathrm{D}_{\mathrm{KL}}(\overline{\mathcal{D}_B} \| \mathcal{U}) \leq \ln(1/\sigma)$. This allows for faster error convergence. Applying these ideas together and setting $\gamma = \sigma/2n$ gives us the following theorem whose proof is deferred to Appendix E.

Theorem 4.2. *For any $(\sigma, 0)$-smooth dataset B of size n, a query class $\mathcal{Q}$ with VC dimension d, $T \in \mathbb{N}$ and $\epsilon > 0$, Smooth Multiplicative Weights Exponential Mechanism is ϵ-differentially private and with probability at least $1 - 2T\,(\gamma/41)^{\mathrm{VCDim}(\mathcal{Q})}$, calculates values v_q for all $q \in \mathcal{Q}$ such that*

$$\max_{q \in \mathcal{Q}} \{|v_q - q(\mathcal{D}_B)|\} \leq \frac{1}{n} + 2\sqrt{\frac{\log(1/\sigma)}{T}} + \frac{10Td\log(2n/\sigma)}{\epsilon n}.$$

Data Release. Above we described a procedure for query answering that relied on the construction of a data set. One could ask whether this leads to a solution to the data release problem as well. An immediate, but ineffective, idea is to output distribution $\overline{\mathcal{D}}$ constructed by our algorithm in the previous section. The problem with this approach is that while $q'(\overline{\mathcal{D}}) \approx q'(\mathcal{D}_B)$ for all queries in the cover $\mathcal{Q}'$, there can be queries $q \in \mathcal{Q} \setminus \mathcal{Q}'$ for which $\left|q(\overline{\mathcal{D}}) - q(\mathcal{D}_B)\right|$ is quite large. This is due to the fact that even though B is $(\sigma, 0)$-smooth (and $\overline{\mathcal{D}_B}$ is σ-smooth), the repeated application of multiplicative update rule may result in distribution $\overline{\mathcal{D}}$ that is far from being smooth.

To address this challenge, we introduce Projected Smooth Multiplicative Weight Exponential Mechanism (Projected Smooth MWEM) that ensures that $\mathcal{D}_t$ is also σ-smooth by projecting it on the convex set of all σ-smooth distributions. More formally, let $\mathcal{K}$ be the polytope of all σ-smooth distributions over $\mathcal{X}$ and let $\tilde{\mathcal{D}}_t$ be the outcome of the multiplicative update rule of **?** at time t. Then, Projected Smooth MWEM mechanism uses $\mathcal{D}_t = \mathrm{argmin}_{\mathcal{D} \in \mathcal{K}} \mathrm{D}_{\mathrm{KL}}(\mathcal{D} \| \tilde{\mathcal{D}}_t)$. To ensure that these projections do not negate the progress made so far, measured by the decrease in KL divergence, we note that for any $\overline{\mathcal{D}_B} \in \mathcal{K}$ and any $\tilde{\mathcal{D}}_t$, we have $\mathrm{D}_{\mathrm{KL}}(\overline{\mathcal{D}_B} \| \tilde{\mathcal{D}}_t) \geq \mathrm{D}_{\mathrm{KL}}(\overline{\mathcal{D}_B} \| \mathcal{D}_t) + \mathrm{D}_{\mathrm{KL}}(\mathcal{D}_t \| \tilde{\mathcal{D}}_t)$. That is, as measured by the decrease in KL divergence, the improvement with respect to $\mathcal{D}_t$ can only be greater than that of $\tilde{\mathcal{D}}_t$. Optimizing parameters T and γ, we obtain the following guarantees. See Appendix F for more details on Projected Smooth MWEM mechanism and its analysis.

Theorem 4.3 (Smooth Data Release)**.** *Let B be a σ-smooth database with n data points. For any $\epsilon, \delta > 0$ and any query set $\mathcal{Q}$ with VC dimension d, Projected Smooth Multiplicative Weight Exponential Mechanism is (ϵ, δ) differentially private and with probability at least $1 - 1/poly\,(n/\sigma)^d$ its outcome $\overline{\mathcal{D}}$ satisfies*

$$\max_{q \in \mathcal{Q}} \left\{\left|q\left(\overline{\mathcal{D}}\right) - q\left(\mathcal{D}_B\right)\right|\right\} \leq O\left(\sqrt{\frac{d}{\epsilon n}\log^{\frac{1}{2}}\left(\frac{1}{\sigma}\right)\log\left(\frac{n}{\sigma}\right)\log\left(\frac{1}{\delta}\right)}\right).$$

5 Conclusions and Open Problems

Our work introduces a framework for smoothed analysis of online and private learning and obtain regret and error bounds that depend only on the VC dimension and the bracketing number of a hypothesis class and are independent of the domain size and Littlestone dimension.

Our work leads to several interesting questions for future work. The first is to characterize learnability in the smoothed setting — via matching lower bounds — in terms of a combinatorial quantity, e.g., bracketing number. In Appendix D, we discuss *sign rank* and its connection to bracketing number as a promising

candidate for this characterization. A related question is whether there are finite VC dimension classes that cannot be learned in presence of smoothed adaptive adversaries.

Let us end this paper by noting that the Littlestone dimension plays a key role in characterizing learnability and algorithm design in the worst-case for several socially and practically important constraints [**??**]. It is essential then to develop models that can bypass Littlestone impossibility results and provide rigorous guidance in achieving these constraints in practical settings.

Acknowledgements

This work was partially supported by the NSF under CCF-1813188, the ARO under W911NF1910294 and a JP Morgan Chase Faculty Fellowship.

A Additional Related Work

Analogous models of smoothed online learning have been explored in prior work. **?** consider online learning when the adversary is constrained in several ways and work with a notion of sequential Rademacher complexity for analyzing the regret. In particular, they study a related notion of smoothed adversary and show that one can learn thresholds with regret of $O(\sqrt{T})$ in presence of smoothed adversaries. **?** consider smoothed online learning in the context online algorithm design. They show that while optimizing parameterized greedy heuristics for Maximum Weight Independent Set imposes linear regret in the worst-case, in presence of smoothing this problem can be learned with sublinear regret (as long they allow per-step runtime that grows with T). **?** consider the same problem with an emphasis on the per-step runtime being logarithmic in T. They show that piecewise constant functions over the interval $[0, 1]$ can be learned efficiently within regret of $O(\sqrt{T})$ against a *non-adaptive* smooth adversary. Our work differs from these by upper bounding the regret using a combinatorial dimension of the hypothesis class and demonstrating techniques that generalize to large class of problems in presence of *adaptive* adversaries.

In another related work, **?** introduce a notion of dispersion in online optimization (where the learner picks an instance and the adversary picks a function) that is a constraint on the number of discontinuities in the adversarial sequence of functions. They show that online optimization can be done efficiently under certain assumptions. Moreover, they show that sequences generated by *non-adaptive* smooth adversaries in one dimension satisfy dispersion. In comparison, our main results in online learning consider the more powerful adaptive adversaries.

Smoothed analysis is also used in a number of other online settings. In the setting of linear contextual bandits, **?** use smoothed analysis to show that the greedy algorithm achieves sublinear regret even though in the worst case it can have linear regret. **?** work in a Bayesian version of the same setting and achieve improved regret bounds for the greedy algorithm. Since several algorithms are known to have sublinear regret in the linear contextual bandit setting even in the worst-case, the main contribution of these papers is to show that the simple and practical greedy algorithm has much better regret guarantees than in the worst-case. In comparison, we work with a setting where no algorithm can achieve sublinear regret in the worst-case.

Smoothed analysis has also been considered in the context of differential privacy. **?** consider differential privacy in the interactive setting, where the queries arrive online. They analyze a multiplicative weights based algorithm whose running time and error they show can be vastly improved in the presence of smoothness. Some of our techniques for query answering and data release are inspired by that line of work. **?** also differential privacy in presence of dispersion and analyze the gaurantees of the exponential mechanism.

Generally, our work is also related to a line of work on online learning in presence of additional assumptions resembling properties exhibited by real life data. **?** consider settings where additional information in terms of an estimator for future instances is available to the learner. They achieve regret bounds that are in terms of the path length of these estimators and can beat $\Omega(\sqrt{T})$ if the estimators are accurate. **?** also considers the importance of incorporating side information in the online learning framework and show that regrets of $O(\log(T))$ in online linear optimization maybe possible when the learner knows a vector that is weakly correlated with the future instances.

More broadly, our work is among a growing line of work on beyond the worst-case analysis of algorithms [**?**] that considers the design and analysis of algorithms on instances that satisfy properties demonstrated

by real-world applications. Examples of this in theoretical machine learning mostly include improved runtime and approximation guarantees of numerous supervised (e.g., [????]), and unsupervised settings (e.g., [??????????]).

B Lack of Uniform Convergence with Adaptive Adversaries

The following example for showing lack of uniform convergence over adaptive sequences is due to **?** and is included here for completeness.

Let $\mathcal{X} = [0,1]$ and $\mathcal{G} = \{g_b(x) = \mathbb{I}(x \geq b) \mid \forall b \in [0,1]\}$ be the set of one-dimensional thresholds. Let the distribution of the noise η_i be the uniform distribution on $(-1/4, 1/4)$. Let $x_1 = 1/2$ and $x_2 = x_3 = \cdots = x_T = 1/4$ if $\eta_1 \leq 0$ while $x_2 = x_3 = \cdots = x_T = 3/4$ otherwise. In this case, we do not achieve concentration for any value of T, as

$$\frac{1}{T}\sum_{t=1}^{T} g_{0.5}(x_t + \eta_t) = \begin{cases} 0 & \text{w.p. } 1/2 \\ 1 & \text{w.p. } 1/2 \end{cases} \qquad \textit{and} \qquad \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} g_{0.5}(x_t + \eta_t)\right] = \frac{1}{2}.$$

C Proofs from Section 3

C.1 Algorithm and its Running Time

While our main focus is to provide sublinear regret bounds for smoothed online learning our analysis also provides an algorithmic solution describe below.

Algorithm 1: Smooth Online Learning

Input: Instance Space $\mathcal{X}$, Hypothesis Class $\mathcal{H}$, Smoothness parmeter σ, Time horizon T

Cover Construction: Compute $\mathcal{H}' \subseteq \mathcal{H}$ that is a γ-cover of $\mathcal{H}$ with respect to the uniform distribution on $\mathcal{X}$ for $\gamma = \frac{\sigma}{4\sqrt{T}}$.

for $t = 1 \dots T$ **do**

  Use a standard online learning algorithm, such as Hedge, on $\mathcal{H}'$ to pick an h_t, where the history of the play is $\{s_\tau\}_{\tau<t}$ and $\{h_\tau\}_{\tau<t}$

  Receive $s_t = (x_t, y_t)$ and suffer loss $\mathrm{err}_{s_t}(h_t)$.

end

The running time of the algorithm comprises of the initial construction of $\mathcal{H}'$ and then running a standard online learning algorithm on $\mathcal{H}'$.

Standard online learning algorithms such as Hedge and FTPL take time polynomial in the size of the cover since in standard implementations they maintain a state corresponding to each hypothesis in $\mathcal{H}'$. In our setting, the size of the cover is $(41\sqrt{T}/\sigma)^d$.

The time required to construct a cover depends on the access we have to the class. One method is to randomly sample a set S with $m = O(\mathrm{VCDim}(\mathcal{H})\, T/\sigma^2)$ points from the domain uniformly and construct all possible labelings on this set induced by the class. The number of labellings of S is bounded by $O(m^{\mathrm{VCDim}(\mathcal{H})})$ by the Sauer–Shelah lemma. The cover is constructed by then finding functions in the class $\mathcal{H}$ that are consistent with each of these labellings. This requires us to be able to find an element in the class consistent with a given labeling, which can be done by a "consistency" oracle. Naively, the above makes 2^m calls to the consistency oracle, one for each possible labeling of S.

The above analysis and runtime can be improved in several ways. First, $\mathcal{H}'$ can be constructed in time $O(m^{\mathrm{VCDim}(\mathcal{H})})$ rather than 2^m. This can be done by constructing the cover in a hierarchical fashion, where the root includes the unlabeled set S and at every level one additional instance in S is labeled by $+1$ or -1. At each node, the consistency oracle will return a function $h \in \mathcal{H}$ that is consistent with the labels so far or state that none exists. Nodes for which no consistent hypothesis so far exists are pruned and will not expand in the next level. Since the total number of leaves is the number of ways in which S can be labeled by $\mathcal{H}$, i.e., $O(m^d)$, the number of calls to the consistency oracle is $O(m^d)$ as well. The runtime of standard online learning algorithms can also be improved significantly when an empirical risk minimization oracle is

available to the learner, in which case a runtime of $O(\sqrt{|\mathcal{H}'|})$ for general classes [**?**] or even polylog$(|\mathcal{H}'|)$) for structured classes [**?**] is possible.

C.2 Proof of Lemma 3.2

At a high level, note that any $f \in \mathcal{F}$ has measure at most ϵ/σ on any (even adaptively chosen) σ-smooth distribution. Therefore, for any fixed f, $\mathbb{E}_{\mathscr{D}}[\sum_{t=1}^{T} f(x_t)] \leq T\epsilon/\sigma$. To achieve this bound over all $f \in \mathcal{F}$, we take a union bound over all such functions.

More formally, for any s

$$\begin{aligned}
\exp\left(s \mathop{\mathbb{E}}_{\mathscr{D}}\left[\max_{f\in\mathcal{F}} \sum_{t=1}^{T} f(x_t)\right]\right) &\leq \mathop{\mathbb{E}}_{\mathscr{D}}\left[\exp\left(s \max_{f\in\mathcal{F}} \sum_{t=1}^{T} f(x_t)\right)\right] && \text{(Jensen's inequqlity)} \\
&\leq \mathop{\mathbb{E}}_{\mathscr{D}}\left[\max_{f\in\mathcal{F}} \exp\left(s \sum_{t=}^{T} f(x_t)\right)\right] && \text{(Monotonicity of exp)} \\
&\leq \sum_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{\mathscr{D}}\left[\exp\left(s \sum_{t=1}^{T} f(x_t)\right)\right]. && (3)
\end{aligned}$$

Consider a fixed $f \in \mathcal{F}$. Note that even when the choice of a σ-smoothed distribution $\mathcal{D}$ depends on earlier realizations of $x_1, \ldots, x_{i-1}$, $\Pr_{x_i\sim\mathcal{D}}[f(x_i)] \leq \frac{\epsilon}{\sigma}$. Therefore, $\sum_{t=1}^{T} f(x_t)$ for $\mathbf{x} \sim \mathscr{D}$ is stochastically dominated by that of a binomial distribution $Bin(T, \epsilon/\sigma)$. Note that $\exp(\cdot)$ is a monotonically increasing functions and let $p = \epsilon/\sigma$. We have

$$\mathop{\mathbb{E}}_{\mathscr{D}}\left[\exp\left(s \sum_{t=1}^{T} f(x_t)\right)\right] \leq \sum_{v=0}^{T} \exp(sv) \binom{T}{v} p^v (1-p)^{T-v} = \left(p(\exp(s)-1)+1\right)^T. \tag{4}$$

Combining Equations (3) and (4) and noting that $\ln(1+x) \leq x$, we have

$$\mathop{\mathbb{E}}_{\mathscr{D}}\left[\max_{f\in\mathcal{F}} \sum_{t=1}^{T} f(x_t)\right] \leq \frac{\ln(|\mathcal{F}|) + Tp\,(\exp(s)-1)}{s}.$$

Let $s = \sqrt{\ln(|\mathcal{F}|)}/Tp$. Note that because $s \in (0,1)$, we have $\exp(s) \leq 1+2s$. Hence, by replacing s in the above inequality we have

$$\mathop{\mathbb{E}}_{\mathscr{D}}\left[\max_{f\in\mathcal{F}} \sum_{t=1}^{T} f(x_t)\right] \in O\left(Tp\sqrt{\ln(|\mathcal{F}|)}\right).$$

C.3 Proof of Theorem 3.3

Consider any hypothesis class $\mathcal{H}'$ and an algorithm that is no-regret with respect to any adaptive adversary on hypotheses in $\mathcal{H}'$. It is not hard to see that

$$\begin{aligned}
\mathbb{E}[\textsc{Regret}(\mathcal{A}, \mathscr{D})] &= \mathop{\mathbb{E}}_{\mathbf{s}\sim\mathscr{D}}\left[\sum_{t=1}^{T} \mathrm{err}_{s_t}(h_t) - \min_{h\in\mathcal{H}} \mathrm{err}_{s_t}(h_t)\right] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{s}\sim\mathscr{D}}\left[\sum_{t=1}^{T} \mathrm{err}_{s_t}(h_t) - \min_{h\in\mathcal{H}'} \sum_{t=1}^{T} \mathrm{err}_{s_t}(h)\right] + \mathop{\mathbb{E}}_{\mathbf{s}\sim\mathscr{D}}\left[\min_{h'\in\mathcal{H}'} \sum_{t=1}^{T} \mathrm{err}_{s_t}(h') - \min \sum_{t=1}^{T} \mathrm{err}_{s_t}(h)\right] \\
&\leq O\left(\sqrt{T\ln(|\mathcal{H}'|)}\right) + \mathop{\mathbb{E}}_{\mathscr{D}}\left[\max_{h\in\mathcal{H}} \min_{h'\in\mathcal{H}'} \sum_{t=1}^{T} 1\left(h(x_t) \neq h'(x_t)\right)\right]. && (5)
\end{aligned}$$

Therefore, it is sufficient to choose an $\mathcal{H}'$ of moderate size such that every function $h \in \mathcal{H}$ has a proxy $h' \in \mathcal{H}'$ even when these functions are evaluated on instances drawn from a *non-iid and adaptive sequence of smooth distributions*. We next describe the choice of $\mathcal{H}'$.

Let $\mathcal{H}'$ be a $\frac{\epsilon}{2}$-net of $\mathcal{H}$ with respect to the uniform distribution $\mathcal{U}$, for an ϵ that we will determine later. Note that any ϵ-bracket with respect to $\mathcal{U}$ is also an ϵ-net, so $|\mathcal{H}'| \leq \mathcal{N}_{[]}(\mathcal{H}, \mathcal{U}, \epsilon/2)$.[3] Let $\mathcal{G}$ be the set of symmetric differences between $h \in \mathcal{H}$ and its closest proxy $h' \in \mathcal{H}'$, that is,

$$\mathcal{G} = \{g_{h,h'}(x) = 1(h(x) \neq h'(x)) \mid \forall h \in \mathcal{H} \text{ and } h' \in \mathcal{H}', \text{ s.t. } \mathop{\mathbb{E}}_{\mathcal{U}}[g_{h,h'}(x)] \leq \epsilon/2\}.$$

Note that because $\mathcal{G}$ is a subset of all the symmetric differences of two functions in $\mathcal{H}$, by Theorem 3.7 its bracketing number is bounded as follows.

$$\mathcal{N}_{[]}(\mathcal{G}, \mathcal{U}, \epsilon/2) \leq \left(\mathcal{N}_{[]}(\mathcal{H}, \mathcal{U}, \epsilon/4)\right)^4. \tag{6}$$

Let $\mathcal{B}(\mathcal{G})$ be the set of upper $\epsilon/2$-brackets of $\mathcal{G}$ with respect to $\mathcal{U}$, i.e., for all $g \in G$, there is $b \in \mathcal{B}(\mathcal{G})$ such that for all $x \in \mathcal{X}$, $g(x) \leq b(x)$ and $\mathbb{E}_{\mathcal{U}}[b(x) - g(x)] \leq \epsilon/2$. Note that

$$\mathop{\mathbb{E}}_{\mathscr{D}}\left[\max_{h \in \mathcal{H}} \min_{h' \in \mathcal{H}'} \sum_{t=1}^{T} 1\left(h(x_t) \neq h'(x_t)\right)\right] = \mathop{\mathbb{E}}_{\mathscr{D}}\left[\max_{g \in \mathcal{G}} \sum_{t=1}^{T} g(x_t)\right] \leq \mathop{\mathbb{E}}_{\mathscr{D}}\left[\max_{b \in \mathcal{B}(\mathcal{G})} \sum_{t=1}^{T} b(x_t)\right],$$

where the last transition is by the fact that $\mathcal{B}(\mathcal{G})$ includes all upper brackets of $\mathcal{G}$.

We now note that $\mathcal{B}(\mathcal{G})$ meets the conditions Lemma 3.2, namely because all $g \in \mathcal{G}$ have measure at most $\epsilon/2$ over $\mathcal{U}$ and $\mathcal{B}(\mathcal{G})$ is the set of $\epsilon/2$-upper brackets of $\mathcal{G}$, we have that $\mathbb{E}_{\mathcal{U}}[b(x)] \leq \epsilon$ for all $b \in \mathcal{B}(\mathcal{G})$. Therefore, by Lemma 3.2 and Equation 6, we have

$$\mathop{\mathbb{E}}_{\mathscr{D}}\left[\max_{b \in \mathcal{B}(\mathcal{G})} \sum_{t=1}^{T} b(x_t)\right] \leq O\left(T\frac{\epsilon}{\sigma}\sqrt{\ln\left(\mathcal{N}_{[]}(\mathcal{H}, \mathcal{U}, \epsilon/4)\right)}\right)$$

Replacing this in Equation 5 we have that

$$\mathbb{E}[\textsc{Regret}(\mathcal{A}, \mathscr{D})] \in O\left(\sqrt{T \ln\left(\mathcal{N}_{[]}(\mathcal{H}, \mathcal{U}, \epsilon/4)\right)} + T\frac{\epsilon}{\sigma}\sqrt{\ln\left(\mathcal{N}_{[]}(\mathcal{H}, \mathcal{U}, \epsilon/4)\right)}\right)$$

Choosing $\epsilon = \sigma/\sqrt{T}$ proves the claim.

C.4 Proof of Theorem 3.6

Consider the map $\psi : \mathcal{X} \to \mathbb{R}^m$ that embeds $\mathcal{G}$ in m dimensions and let $\mathcal{H}$ be the class of halfspaces in $\mathbb{R}^m$. We want to bound the bracketing number of $\mathcal{G}$ by that of $\mathcal{H}$. Let $\mathcal{B}(\mathcal{H}) = \{[h_i, h^i]\}_i$ be an ϵ-bracketing for $\mathcal{H}$ with respect to a measure μ that we will specify later. Consider the set of brackets $\mathcal{B}' = \{[h_i \circ \psi, h^i \circ \psi] \mid \text{ for all } [h_i, h^i] \in \mathcal{B}(\mathcal{H})\}$. We first argue that $\mathcal{B}'$ is a bracketing for $\mathcal{G}$ with respect to ν. To see this, note that any $g \in \mathcal{G}$ can be expressed as $g = h \circ \psi$ for some halfspace h. Considering the bracket $[h_i, h^i] \ni h$ in $\mathcal{B}(\mathcal{H})$. Note that $h_i \circ \psi \preceq h \circ \psi \preceq h^i \circ \psi$ and thus $g \in [h_i \circ \psi, h^i \circ \psi]$. We next argue that these are ϵ-brackets under measure ν. Let μ be the measure such that to sample $z \sim \mu$ we first sample $x \sim \nu$ and let $z = \psi(x)$. Note that

$$\Pr_{x \sim \nu}\left[h^i(\psi(x)) \neq h_i(\psi(x))\right] = \Pr_{z \sim \mu}\left[h^i(z) \neq h_i(z)\right] \leq \epsilon,$$

where the last transition is by the fact that $\mathcal{B}(\mathcal{H})$ is an ϵ-bracketing for $\mathcal{H}$ with respect to μ. This concludes that $\mathcal{B}'$ is an ϵ-bracketing for $\mathcal{G}$ with respect to ν. We complete the proof by using Theorem 3.4 to bound $|\mathcal{B}'| = |\mathcal{B}(\mathcal{H})| \leq (m/\epsilon)^{O(m)}$.

C.5 Proof of Theorem 3.7

We first consider the case of $k = 2$ and then extend our argument to general k. Let $\epsilon' = \epsilon/k$ and let $\mathcal{B}(\mathcal{F}_1)$ and $\mathcal{B}(\mathcal{F}_2)$ be ϵ'-bracketings for $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively.

[3] Alternatively, we can bound $|\mathcal{H}'| \leq (41/\epsilon)^{\text{VCDim}(\mathcal{H})}$ by **?**.

For $\mathcal{F}_1 \cdot \mathcal{F}_2$, construct $\mathcal{B} = \{[f_\ell \cap g_\ell, f^u \cap g^u] \mid \text{for all } [f_\ell, f^u] \in \mathcal{B}(\mathcal{F}_1) \text{ and } [g_\ell, g^u] \in \mathcal{B}(\mathcal{F}_2)\}$. First note for any $f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2$, $f_1 \cap f_2$ is included in one of these brackets. In particular, for brackets $[f_\ell, f^u] \ni f_1$ and $[g_\ell, g^u] \ni f_2$, we have that $f_\ell \cap g_\ell \preceq f_1 \cap f_2 \preceq f^u \cap g^u$ and $[f_\ell \cap g_\ell, f^u \cap g^u] \in \mathcal{B}$. Furthermore,

$$\begin{aligned}\Pr_{x\sim\mu}\left[(f_\ell(x) \cap g_\ell(x)) \neq (f^u(x) \cap g^u(x))\right] &\leq \Pr_{x\sim\mu}\left[(f_\ell(x) \cap g_\ell(x)) \neq (f_\ell(x) \cap g^u(x))\right] \\ &\quad + \Pr_{x\sim\mu}\left[(f_\ell(x) \cap g^u(x)) \neq (f^u(x) \cap g^u(x))\right] \\ &\leq 2\epsilon'.\end{aligned}$$

Therefore, $\mathcal{B}$ is a $2\epsilon'$-bracketing for $\mathcal{F}_1 \cdot \mathcal{F}_2$ of size $\mathcal{N}_{[]}(\mathcal{F}_1, \mu, \epsilon') \cdot \mathcal{N}_{[]}(\mathcal{F}_2, \mu, \epsilon')$. Repeating this inductively and using $\epsilon' = \epsilon/k$, we get the claim for k classes.

Similarly, for $\mathcal{F}_1 + \mathcal{F}_2$, construct $\mathcal{B} = \{[f_\ell \cup g_\ell, f^u \cup g^u] \mid \text{for all } [f_\ell, f^u] \in \mathcal{B}(\mathcal{F}_1) \text{ and } [g_\ell, g^u] \in \mathcal{B}(\mathcal{F}_2)\}$. First note for any $f_1 \in \mathcal{F}$ and $f_2 \in \mathcal{F}_1$ and their respective brackets $[f_\ell, f^u] \ni f_1$ and $[g_\ell, g^u] \ni f_2$, we have that $f_\ell \cup g_\ell \preceq f_1 \cup f_2 \preceq f^u \cup g^u$ and $[f_\ell \cup g_\ell, f^u \cup g^u] \in \mathcal{B}$. Furthermore,

$$\begin{aligned}\Pr_{x\sim\mu}\left[(f_\ell(x) \cup g_\ell(x)) \neq (f^u(x) \cup g^u(x))\right] &\leq \Pr_{x\sim\mu}\left[f_\ell(x) \neq f^u(x)\right] + \Pr_{x\sim\mu}\left[g_\ell(x) \neq g^u(x)\right] \\ &\leq 2\epsilon'.\end{aligned}$$

Therefore, $\mathcal{B}$ is a $2\epsilon'$-bracketing for $\mathcal{F}_1 + \mathcal{F}_2$ of size $\mathcal{N}_{[]}(\mathcal{F}_1, \mu, \epsilon') \cdot \mathcal{N}_{[]}(\mathcal{F}_2, \mu, \epsilon')$. Repeating this inductively and using $\epsilon' = \epsilon/k$, we get the claim for k classes.

As for the $\mathcal{G}$, the set of all symmetric differences, note that $f_1 \Delta f_2 = (f_1 \cup f_2) \setminus (f_1 \cap f_2) = (f_1 \cup f_2) \cap \overline{(f_1 \cap f_2)}$. Furthermore, for any class $\mathcal{F}$, the class $\overline{\mathcal{F}} = \{\overline{f} \mid \forall f \in \mathcal{F}\}$ has the same bracketing number as $\mathcal{F}$. Therefore, the bracketing number of $\mathcal{G}$ follows from using the bracketing number $\mathcal{F} + \mathcal{F}$, $\overline{\mathcal{F} + \mathcal{F}}$, and their intersection.

C.6 Proof of Corollary 3.8

The set of polynomial threshold functions in n variables and of degree d is embeddable as halfspaces in $O(n^d)$ dimensions using the map

$$\phi(x_1, \ldots, x_n) = \left(\prod_{i \in S} x_i\right)_{S \in \{1,\ldots,n\}^{\leq d}},$$

which maps variables to all monomial of degree d. It can be seen that the number of monomials of degree at most d in n variables is given by $\binom{n+d+1}{d+1}$ which is approximately $O\left(n^d\right)$ when d is small. Combining Theorem 3.6 and Theorem 3.4 completes the proof for polynomial threshold functions.

A k-polytope in $\mathbb{R}^n$ is an intersection of k-halfspaces in $\mathbb{R}^n$. Combining Theorem 3.7 and Theorem 3.4 completes the proof.

D More Details on Bracketing Number and Sign Rank

Though bracketing numbers are a fundamental concept in statistics, until recently their connection to VC theory was not well understood. **??** show that for countable (can be generalized to classes that are well approximated by countable classes) classes with finite VC dimension the bracketing numbers with respect to any measure is finite (this establishes what is known as a universal Gilvenko–Cantelli theorem under ergodic sampling.)

Theorem D.1 (Finite Bracketing Bounds for VC Classes)**.** *Let $\mathcal{C}$ be a countable class with finite VC dimension. Then, $\mathcal{N}_{[]}(\mathcal{C}, \mu, \epsilon) < \infty$.*

Though the above theorem proves that ϵ-bracketing numbers are finite, their growth rate in $1/\epsilon$ can be arbitrarily large. See **?** for some interesting examples of classes where the bracketing numbers grow arbitrarily fast.

Another combinatorial quantity that can help bound the regret in presence of adaptive smooth adversaries is *sign rank*.

Definition D.2 (Sign Rank)**.** *Let $\mathcal{X}$ be an instance space and let $\mathcal{F}$ be a class. We can denote the class naturally as $\{-1,1\}$-valued $\mathcal{X}\times\mathcal{F}$ matrix $M_{\mathcal{F}}$ where the entry corresponding to (x,f) is $f(x)$. The sign rank of a class is the highest rank of a real matrix that agrees with a finite submatrix of $M_{\mathcal{F}}$ in sign. If this is unbounded, the class is said to have infinite sign rank.*

The sign rank of a class captures the dimension in which the class can be embedded as thresholds.

Fact D.3 (Sign Rank Embedding, see e.g. **?**)**.** *The sign rank of a class corresponds to the smallest dimension d that the class can be embedded as thresholds.*

Theorem 3.6 effectively says that classes with small sign rank have a slowly growing bracketing numbers and thus have low regret in the smoothed online learning setting. Thus, the complexity of smoothed online learning lies somewhere in between the sign rank and VC dimension. On the other hand, it is known that even classes with small VC dimension can have arbitrarily large sign rank [**???**]. An intermediate question is whether classes with slow growing bracketing number also have good sign rank. It would be interesting to characterize the complexity of smoothed online learning in terms of either the sign rank or bracketing numbers.

E Query Answering

E.1 Smooth MWEM Algorithm

Algorithm 2: Smooth Multiplicative Weights Exponential Mechanism

Input: Universe $\mathcal{X}$ with $|\mathcal{X}|=N$, Data set B with n records, Query set $\mathcal{Q}$, Privacy parameters ϵ and δ, Smoothness parameter σ.

Let $\mathcal{D}_0(x) = 1/N$ for all $x\in\mathcal{X}$.
Cover Construction: Compute $\mathcal{Q}'\subseteq\mathcal{Q}$ that is a γ-cover of $\mathcal{Q}$ with respect to the uniform distribution for $\gamma=\frac{\sigma}{2n}$.
for $i=1\dots T$ **do**
Exponential Mechanism: Sample $q_i\in\mathcal{Q}'$ according to the exponential mechanism with parameter $\epsilon/2T$ and score function

$$s_i(\mathcal{D}_B,q)=n\left|q\left(\mathcal{D}_{i-1}\right)-q(\mathcal{D}_B)\right|.$$

Laplace Mechanism: Let $m_i=q_i\left(\mathcal{D}_B\right)+\frac{1}{n}Lap\left(2T/\epsilon\right)$.
Multiplicative Update: Update $\mathcal{D}_{i-1}$ using the rule

$$\mathcal{D}_i(x)\propto\mathcal{D}_{i-1}(x)\exp\left(\frac{q_i(x)\left(m_i-q_i(\mathcal{D}_{i-1})\right)}{2}\right).$$

end
Let $\overline{\mathcal{D}}=\frac{1}{T}\sum_{i=1}^{T}\mathcal{D}_i$.
Output: For each $q\in\mathcal{Q}$, answer with $v_q=q'\left(\overline{\mathcal{D}}\right)$ where q' is the closest function in $\mathcal{Q}'$ to q.

E.2 Proof of Theorem 4.2

In this section we prove the following theorem.

Theorem 4.2 (restated). *For any $(\sigma,0)$-smooth dataset B of size n, a query class $\mathcal{Q}$ with VC dimension d, $T\in\mathbb{N}$ and $\epsilon>0$, Smooth Multiplicative Weights Exponential Mechanism is ϵ-differentially private and with probability at least $1-2T\left(\gamma/41\right)^{\mathrm{VCDim}(\mathcal{Q})}$, calculates values v_q for all $q\in\mathcal{Q}$ such that*

$$\max_{q\in\mathcal{Q}}\left\{\left|v_q-q\left(\mathcal{D}_B\right)\right|\right\}\le\frac{1}{n}+2\sqrt{\frac{\log\left(1/\sigma\right)}{T}}+\frac{10Td\log\left(2n/\sigma\right)}{\epsilon n}.$$

Let us first provide a few useful lemmas.

Lemma E.1 (Cover under Smoothness)**.** *Let B be $(\sigma, 0)$-smooth data set. Let $\mathcal{Q}' \subseteq \mathcal{Q}$ be a γ-cover of $\mathcal{Q}$ under the uniform distribution. For a $q \in \mathcal{Q}$, let $q' \in \mathcal{Q}$ be such that $\Pr_{x\sim\mathcal{U}}[q(x) \neq q'(x)] \leq \gamma$. Then,*

$$|q(\mathcal{D}_B) - q'(\mathcal{D}_B)| \leq \frac{2\gamma}{\sigma}.$$

Proof. From the $(\sigma, 0)$-smoothness of B, we get

$$\begin{aligned}
|q(\mathcal{D}_B) - q'(\mathcal{D}_B)| &= \left|q\left(\overline{\mathcal{D}_B}\right) - q'\left(\overline{\mathcal{D}_B}\right)\right| \\
&\leq \sum_{x\in D} |(q(x) - q'(x))|\, \overline{\mathcal{D}_B}(x) \\
&\leq \sum_{x\in\mathcal{X}} 2\mathbb{I}(q(x) \neq q'(x))\, \overline{\mathcal{D}_B}(x) \\
&\leq \frac{2}{\sigma} \sum_{x\in\mathcal{X}} \mathbb{I}(q(x) \neq q'(x))\, \mathcal{U}(x) \\
&\leq \frac{2}{\sigma} \Pr_{x\sim\mathcal{U}}[q(x) \neq q'(x)] \\
&\leq \frac{2\gamma}{\sigma}
\end{aligned}$$

as required. □

Define the potential function $\Psi_i = \sum_{x\in\mathcal{X}} \overline{\mathcal{D}_B}(x) \log\left(\overline{\mathcal{D}_B}(x)/\mathcal{D}_i(x)\right)$, where $\overline{\mathcal{D}_B}$ is a corresponding σ-smooth distribution that matches the query answers for the $(\sigma, 0)$-smooth data set B. Here we make a few observations about the potential function.

Fact E.2. *For all $i \leq T$, we have $\Psi_i \geq 0$. Furthermore, $\Psi_0 \leq \log\frac{1}{\sigma}$. As a result, $\Psi_0 - \Psi_T \leq \log\frac{1}{\sigma}$.*

Proof. The first claim follows from the positivity of the KL divergence. For the second one, recall that from the σ-smoothness of $\mathcal{D}_B$ and the fact that $\mathcal{D}_1$ is the uniform distribution, we have $\mathcal{D}_B(x) \leq \sigma^{-1}\mathcal{D}_0(x)$ for all $x \in \mathcal{X}$.

$$\Psi_0 = \sum_{x\in\mathcal{X}} \overline{\mathcal{D}_B}(x) \log\frac{\overline{\mathcal{D}_B}(x)}{\mathcal{D}_0(x)} \leq \sum_{x\in\mathcal{X}} \overline{\mathcal{D}_B}(x) \log\frac{1}{\sigma} = \log\frac{1}{\sigma}$$

as required. □

Below is a direct adaptation of a result of **?** for bounding the change in the potential functions.

Lemma E.3 (Lemma A.4 in **?**)**.**

$$\Psi_{i-1} - \Psi_i \geq \left(\frac{q_i(\mathcal{D}_{i-1}) - q_i(\overline{\mathcal{D}_B})}{2}\right)^2 - \left(\frac{m_i - q_i(\overline{\mathcal{D}_B})}{2}\right)^2.$$

Lemma E.4 (Exponential and Laplace Mechanism guarantees)**.** *With probability at least $1 - {}^{2T}/_{|\mathcal{Q}'|}$, we have*

$$|q_i(\mathcal{D}_{i-1}) - q_i(\mathcal{D}_B)| \geq \max_{q'\in\mathcal{Q}'} \{q'(\mathcal{D}_i) - q'(\mathcal{D}_B)\} - \frac{8T\log|\mathcal{Q}'|}{\epsilon n}$$

and

$$|m_i - q_i(\mathcal{D}_B)| \leq \frac{2T\log|\mathcal{Q}'|}{\epsilon n}.$$

Here we recall again the error guarantees from **?**.

Theorem E.5 (**?**). *For any data set B of size n, a finite query class $\mathcal{Q}$, $T \in \mathbb{N}$ and $\epsilon > 0$, MWEM is ϵ-differentially private and with probability at least $1 - {}^{2T}/_{|\mathcal{Q}|}$ produces a distribution $\overline{\mathcal{D}}$ over $\mathcal{X}$ such that*

$$\max_{q \in \mathcal{Q}} \left\{ \left| q\left(\overline{\mathcal{D}}\right) - q\left(\mathcal{D}_B\right) \right| \right\} \leq 2\sqrt{\frac{\log|\mathcal{X}|}{T}} + \frac{10T\log|\mathcal{Q}|}{\epsilon n}.$$

Proof of Theorem 4.2. Our proof closely resembles that of Theorem E.5 from **?**. Note that since B is $(\sigma, 0)$-smooth, we have a σ-smooth distribution $\overline{\mathcal{D}_B}$ with $\overline{\mathcal{D}_B}(x) \leq \frac{1}{\sigma N}$ such that for all $q \in \mathcal{Q}$, $q\left(\mathcal{D}_B\right) = q\left(\overline{\mathcal{D}_B}\right)$. Furthermore, note that we chose a cover $\mathcal{Q}' \subseteq \mathcal{Q}$. Therefore, $q'\left(\mathcal{D}_B\right) = q'\left(\overline{\mathcal{D}_B}\right)$ holds for all $q' \in \mathcal{Q}'$ as well.

Note that since $q'\left(\mathcal{D}_B\right) = q'\left(\overline{\mathcal{D}_B}\right)$ for all $q' \in \mathcal{Q}'$, we can replace this in the above equation. For the sake of completeness, we sketch the rest of the proof. From Jensen's inequality, we have

$$\max_{q' \in \mathcal{Q}'} \left| q'\left(\overline{\mathcal{D}}\right) - q'\left(\mathcal{D}_B\right) \right| \leq \frac{1}{T}\sum_{i=1}^{T} \max_{q' \in \mathcal{Q}'} \left| q'\left(\mathcal{D}_i\right) - q'\left(\mathcal{D}_B\right) \right|. \tag{7}$$

From Lemma E.4 and Lemma E.3, we get that with probability at least $1 - 2T/|\mathcal{Q}'|$, we get

$$\Psi_{i-1} - \Psi_i \geq \left(\frac{\max_{q' \in \mathcal{Q}'} \left\{ q'\left(\mathcal{D}_i\right) - q'\left(\mathcal{D}_B\right) \right\} - \frac{8T\log|\mathcal{Q}'|}{\epsilon n}}{2} \right)^2 - \left(\frac{T\log|\mathcal{Q}|}{\epsilon n} \right)^2.$$

Rearranging this and taking the average, we get

$$\frac{1}{T}\sum_{i=1}^{T} \max_{q' \in \mathcal{Q}'} \left| q'\left(\mathcal{D}_i\right) - q'\left(\mathcal{D}_B\right) \right| \leq \frac{1}{T}\sum_{i=1}^{T} \left[\sqrt{4\left(\Psi_{i-1} - \Psi_i\right) + \frac{4T^2\log^2|\mathcal{Q}'|}{n^2\epsilon^2}} + \frac{8T\log|\mathcal{Q}'|}{n\epsilon} \right].$$

Applying the concavity of the square root function i.e., $\frac{1}{T}\sum_{i=1}^{T}(x_i)^{1/2} \leq \left(\frac{1}{T}\sum_{i=1}^{T} x_i\right)^{1/2}$,

$$\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T} \max_{q \in \mathcal{Q}'} \left| q'\left(\mathcal{D}_i\right) - q'\left(\mathcal{D}_B\right) \right| &\leq \sqrt{\sum_{i=1}^{T} \frac{4\left(\Psi_{i-1} - \Psi_i\right)}{T} + \frac{4T^2\log^2|\mathcal{Q}'|}{n^2\epsilon^2}} + \frac{8T\log|\mathcal{Q}'|}{n\epsilon} \\
&\leq \sqrt{\frac{4\left(\Psi_0 - \Psi_T\right)}{T} + \frac{4T^2\log^2|\mathcal{Q}'|}{n^2\epsilon^2}} + \frac{8T\log|\mathcal{Q}'|}{n\epsilon} \\
&\leq \sqrt{\frac{4\log\left(\frac{1}{\sigma}\right)}{T} + \frac{4T^2\log^2|\mathcal{Q}'|}{n^2\epsilon^2}} + \frac{8T\log|\mathcal{Q}'|}{n\epsilon} \\
&\leq 2\sqrt{\frac{\log\left(\frac{1}{\sigma}\right)}{T}} + \frac{10T\log|\mathcal{Q}'|}{n\epsilon}.
\end{aligned}$$

The second inequality follows by summing the telescoping series. The third follows from Fact E.2. The last equation follows from the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for all positive x, y. Using Equation 7 and the fact that $|\mathcal{Q}|' \leq \left({}^{41}/_{\gamma}\right)^d$ we have

$$\max_{q' \in \mathcal{Q}'} \left| q'\left(\overline{\mathcal{D}}\right) - q'\left(\mathcal{D}_B\right) \right| \leq 2\sqrt{\frac{\log\left({}^{1}/_{\sigma}\right)}{T}} + \frac{10Td\log\left({}^{2n}/_{\sigma}\right)}{\epsilon n}.$$

Let $v_q = q'(\overline{\mathcal{D}})$ for $q' \in \mathcal{Q}'$ that is the closest hypothesis to q with respect to the uniform distribution. Then

$$\begin{aligned}
\left| q\left(\mathcal{D}_B\right) - v_q \right| &= \left| q\left(\mathcal{D}_B\right) - q'\left(\mathcal{D}_B\right) + q'\left(\mathcal{D}_B\right) - q'\left(\overline{\mathcal{D}}\right) \right| \\
&\leq \left| q\left(\mathcal{D}_B\right) - q'\left(\mathcal{D}_B\right) \right| + \left| q'\left(\mathcal{D}_B\right) - q'\left(\overline{\mathcal{D}}\right) \right| \\
&\leq \frac{2\gamma}{\sigma} + 2\sqrt{\frac{\log {}^{1}/_{\sigma}}{T}} + \frac{10Td\log\left({}^{41}/_{\gamma}\right)}{\epsilon n}.
\end{aligned}$$

Setting $\gamma = \frac{\sigma}{4n}$, we get the desired result. □

Setting $T = \epsilon^{2/3} n^{2/3} \log^{1/3}(1/\sigma)\, d^{-2/3} \log^{-2/3}(2n/\sigma)$, we get $(\epsilon, 0)$ differential privacy with

$$\max_{q \in \mathcal{Q}} \{|v_q - q(\mathcal{D}_B)|\} \leq O\left(\sqrt[3]{\frac{d \log(1/\sigma) \log(2n/\sigma)}{n\epsilon}}\right).$$

Also, as noted in **?**, one can use adaptive k-fold composition (see e.g. **?**) to get (ϵ, δ)-differential privacy with

$$\max_{q \in \mathcal{Q}} \{|v_q - q(\mathcal{D}_B)|\} \leq O\left(\sqrt{\frac{d}{\epsilon n} \log^{\frac{1}{2}}\left(\frac{1}{\sigma}\right) \log\left(\frac{n}{\sigma}\right) \log\left(\frac{1}{\delta}\right)}\right).$$

E.3 Running Time of the Algorithm

The running time of the algorithm is similar to the running time of the MWEM algorithm of **?**. The main additional step is the construction of the cover $\mathcal{Q}'$. Similar to Appendix C.1 , this cover can be constructed in time $O(|\mathcal{Q}'|)$. The exponential mechanism requires $O\left(n|\mathcal{Q}|'\right)$ to evaluate all the queries on the cover and time $O\left(|\mathcal{Q}|'|\mathcal{X}|\right)$ to execute each iteration of the algorithm. Recall that $|\mathcal{Q}'| \leq (41n/\sigma)^d$, thus the running time is bounded by $O\left(n(41n/\sigma)^d + T(41n/\sigma)^d |\mathcal{X}|\right)$.

This runtime can also be improved using several theoretical tricks, e.g., $q(\mathcal{D}_i)$ can be approximated by taking random points from $\mathcal{D}_i$ in time that is independent of $\mathcal{X}$.

Note that the runtime of our algorithm improves upon the runtime of MWEM by using smaller query sets. As noted in **?**, their algorithm is amenable to many optimizations and modifications that make it very fast and practical **?**.

F Data Release

F.1 Projected Smooth MWEM Algorithm

Algorithm 3: Projected Smooth Multiplicative Weight Exponential Mechanism

Input: Universe $\mathcal{X}$ with $|\mathcal{X}| = N$, Data set B with n records, Query set $\mathcal{Q}$, Privacy parameters ϵ and δ, Smoothness parameter σ.

Let $\mathcal{D}_0(x) = 1/N$ for all $x \in \mathcal{X}$.

Cover Construction: Compute $\mathcal{Q}' \subseteq \mathcal{Q}$ that is a γ-cover of $\mathcal{Q}$ with respect to the uniform distribution for $\gamma = \frac{\sigma}{2n}$.

for $i = 1 \dots T$ **do**

Exponential Mechanism: Sample $q_i \in \mathcal{Q}'$ according to the exponential mechanism with parameter $\epsilon/2T$ and score function

$$s_i(\mathcal{D}_B, q) = n\left|q\left(\mathcal{D}_{i-1}\right) - q(\mathcal{D}_B)\right|.$$

Laplace Mechanism: Let $m_i = q_i\left(\mathcal{D}_B\right) + \frac{1}{n} Lap\left(2T/\epsilon\right)$.

Multiplicative Update: Update $\mathcal{D}_{i-1}$ using the rule

$$\tilde{\mathcal{D}}_i(x) \propto \mathcal{D}_{i-1}(x) \exp\left(\frac{q_i(x)\left(m_i - q_i(\mathcal{D}_{i-1})\right)}{2}\right).$$

KL Projection: Project $\tilde{\mathcal{D}}_i$ onto the polytope $\mathcal{K} = \left\{\mathbf{z} : z_i \geq 0,\ \sum_{i=1}^{N} z_i = 1, z_i \leq \frac{1}{\sigma N}\right\}$ of smooth distributions:

$$\mathcal{D}_i = \underset{\mathcal{D} \in \mathcal{K}}{\operatorname{argmin}}\, \mathrm{D}_{\mathrm{KL}}(\mathcal{D} \| \tilde{\mathcal{D}}_i)$$

end

Let $\overline{\mathcal{D}} = \frac{1}{T}\sum_{i=1}^{T} \mathcal{D}_i$.

Output: Distribution $\overline{\mathcal{D}}$.

F.2 Proof of Theorem 4.3

As before, let $\overline{\mathcal{D}_B}$ be a corresponding σ-smooth distribution that matches the query answers for the $(\sigma, 0)$-smooth data set B. Define $\Psi_i = \sum_{x \in \mathcal{X}} \overline{\mathcal{D}_B}(x) \log\left(\overline{\mathcal{D}_B}(x)/\mathcal{D}_i(x)\right)$ and $\tilde{\Psi}_i = \sum_{x \in \mathcal{X}} \overline{\mathcal{D}_B}(x) \log\left(\overline{\mathcal{D}_B}(x)/\tilde{\mathcal{D}}_i(x)\right)$ as the intermediate potential. From Lemma E.3, we know

$$\Psi_{i-1} - \tilde{\Psi}_i \geq \left(\frac{q_i\left(\mathcal{D}_{i-1}\right) - q_i(\mathcal{D}_B)}{2}\right)^2 - \left(\frac{m_i - q_i(\mathcal{D}_B)}{2}\right)^2.$$

Using the properties of relative entropy, we show the following claim.

Claim F.1. *For every $i \leq T$, we have $\tilde{\Psi}_i \geq \Psi_i$.*

Proof. The claim follows from the following fact about the KL divergence. Let

$$\mathcal{D}_i = \underset{\mathcal{D} \in \mathcal{K}}{\operatorname{argmin}}\, \mathrm{D}_{\mathrm{KL}}(\mathcal{D} \| \tilde{\mathcal{D}}_i)$$

for some convex set $\mathcal{K}$. Then, for $\overline{\mathcal{D}_B} \in \mathcal{K}$,

$$\mathrm{D}_{\mathrm{KL}}(\overline{\mathcal{D}_B} \| \tilde{\mathcal{D}}_i) \geq \mathrm{D}_{\mathrm{KL}}\left(\overline{\mathcal{D}_B} \| \mathcal{D}_i\right) + \mathrm{D}_{\mathrm{KL}}\left(\mathcal{D}_i \| \tilde{\mathcal{D}}_i\right).$$

The claim follows by $\tilde{\Psi}_i = \mathrm{D}_{\mathrm{KL}}(\mathcal{D}_B \| \tilde{\mathcal{D}}_i)$, $\Psi_i = \mathrm{D}_{\mathrm{KL}}\left(\mathcal{D}_B \| \mathcal{D}_i\right)$ and $\mathrm{D}_{\mathrm{KL}}\left(\mathcal{D}_i \| \tilde{\mathcal{D}}_i\right) \geq 0$. □

Together this gives

$$\Psi_{i-1} - \Psi_i \geq \left(\frac{q_i\left(\mathcal{D}_{i-1}\right) - q_i(\mathcal{D}_B)}{2}\right)^2 - \left(\frac{m_i - q_i(\mathcal{D}_B)}{2}\right)^2.$$

The remainder of the analysis follows that of Theorem 4.2. Note that we have $\overline{\mathcal{D}}$ is σ-smooth since each $\mathcal{D}_i \in \mathcal{K}$ and $\mathcal{K}$ is a convex set. By Lemma E.1, we have $\left|q'\left(\overline{\mathcal{D}}\right) - q\left(\overline{\mathcal{D}}\right)\right| \leq {}^{2\gamma}/_{\sigma}$. Thus,

$$\begin{aligned}
\left|q\left(\mathcal{D}_B\right) - q\left(\overline{\mathcal{D}}\right)\right| &= \left|q\left(\mathcal{D}_B\right) - q'\left(\mathcal{D}_B\right) + q'\left(\mathcal{D}_B\right) - q'\left(\overline{\mathcal{D}}\right) + q'\left(\overline{\mathcal{D}}\right) - q\left(\overline{\mathcal{D}}\right)\right| \\
&\leq \left|q\left(\mathcal{D}_B\right) - q'\left(\mathcal{D}_B\right)\right| + \left|q'\left(\mathcal{D}_B\right) - q'\left(\overline{\mathcal{D}}\right)\right| + \left|q'\left(\overline{\mathcal{D}}\right) - q\left(\overline{\mathcal{D}}\right)\right| \\
&\leq \frac{4\gamma}{\sigma} + 2\sqrt{\frac{\log {}^{1}/_{\sigma}}{T}} + \frac{10Td\log\left({}^{41}/_{\gamma}\right)}{\epsilon n}.
\end{aligned}$$

Setting $\gamma = {}^{\sigma}/_{4n}$, we get

$$\left|q\left(\mathcal{D}_B\right) - q\left(\overline{\mathcal{D}}\right)\right| = \frac{1}{n} + 2\sqrt{\frac{\log\left({}^{1}/_{\sigma}\right)}{T}} + \frac{10Td\log\left({}^{4n}/_{\sigma}\right)}{\epsilon n}.$$

F.3 Running Time of Projected Smooth Multiplicative Weights Exponential Mechanism

The running time is similar to the running time Smooth Multiplicative Weights Exponential Mechanism,with the additional projection step in each step. Note that the projection in each step is a convex program and can be solved in time poly $(|\mathcal{X}|)$. This gives us a total running time of $O\left(n\left(41n/\sigma\right)^d + T\left(41n/\sigma\right)^d|\mathcal{X}| + T\text{poly}(|\mathcal{X}|)\right)$.

In addition to the improvements discussed in the previous sections, the projection step can be performed faster by taking an approximate Bregman projection as considered by **?**. Incorporating this into our algorithm would lead to significant speed ups.

G Smooth Data Release using SmallDB Algorithm

In this section,, we look at a different algorithm to get differential privacy when dealing with (σ, χ)-smooth data sets. Our algorithm displayed below uses several pieces that have been introduced by **?** and **?**.

Algorithm 4: Subsampled Net Mechanism

Input: Database B of size n, Query set $\mathcal{Q}$, Privacy parameter ϵ, Subsampling parameter M, Accuracy parameter γ.

Sample (with replacement) a subset V of size M from $\mathcal{X}$.

Sample B' from amongst all data sets supported on V of size

$$O\left(\frac{d}{\gamma^2}\right)$$

with probability proportional to

$$\exp\left(-\frac{\epsilon \cdot n \cdot s\left(\mathcal{D}_{B'}, \mathcal{D}_B\right)}{2}\right)$$

where $s\left(\mathcal{D}_{B'}, \mathcal{D}_B\right) = \max_{q\in\mathcal{Q}}|q(\mathcal{D}_B) - q(\mathcal{D}_{B'})|$.

Output: Database B'

First, we analyze the privacy of this algorithm.

Theorem G.1. *The Subsampled Net Mechanism is* $(\epsilon, 0)$ *differentially private.*

Proof. The privacy claim follows from the privacy of the exponential mechanism. □

Next we bound the error of this mechanism. Let us recall the standard uniform convergence bound.

Fact G.2 (Uniform Convergence for VC Classes, see e.g. **?**)**.** *Let $\mathcal{X}$ be the domain, $\mathcal{Q}$ be a class of queries over $\mathcal{X}$ with VC dimension d and let $\mathcal{D}$ be a distribution. Let $\mathcal{D}'$ be a distribution gotten by sampling $O\left((\log(2/\eta)+d)/\gamma^2\right)$ items iid from $\mathcal{D}$ and normalizing the frequencies. Then, with probability $1-\eta$, for all $q \in \mathcal{Q}$, $|q(\mathcal{D}') - q(\mathcal{D})| \leq \gamma$.*

In the following, we use the above fact to show that a randomly sampled subset of the universe approximates a (σ, χ)-smooth database. The proof largely follows the domain reduction lemma of **?** that achieve a similar bond by with a dependence on $\log(|\mathcal{Q}|)$. We include this proof for completeness.

Lemma G.3. *Let $\mathcal{X}$ be a data universe and $\mathcal{Q}$ a collection of queries over $\mathcal{X}$ with VC dimension d and $\mathcal{D}$ be (σ, χ)-smooth with respect to $\mathcal{Q}$. Let $V \subset \mathcal{X}$ of size M be sampled from $\mathcal{X}$ at random with replacement with*

$$M = O\left(\frac{\log(1/\eta) + d}{\sigma\gamma^2}\right).$$

Then, with probability $1-\eta$, there exists a $\mathcal{D}'$ on V such that for all $q \in \mathcal{Q}$

$$|q(\mathcal{D}) - q(\mathcal{D}')| \leq \chi + \gamma.$$

Proof. Let $\mathcal{D}_1$ be σ-smooth distribution that witnesses the (σ, χ)-smoothness of $\mathcal{D}$. If we could sample from $\mathcal{D}_1$, we would be done from Fact G.2. But we want to get a subset that is oblivious to the distribution $\mathcal{D}$. To achieve this, we use the smoothness of $\mathcal{D}_1$.

The idea is to sample from $\mathcal{D}_1$ using rejection sampling. Since $\mathcal{D}_1$ is σ-smooth, the following procedure produces samples from $\mathcal{D}_1$: sample from the uniform distribution and accept sample u with probability $\sigma N \mathcal{D}_1(u)$. Note that accepted samples are distributed according to $\mathcal{D}_1$. We repeat this process until $O\left((\log(2/\eta)+d)/\gamma^2\right)$ samples are accepted. Since the accepted samples are distributed according to $\mathcal{D}_1$, from Fact G.2, there is a distribution $\mathcal{D}_2$ supported on the accepted samples such that with probability at least $1-\eta/2$ for all $q \in \mathcal{Q}$,

$$|q(\mathcal{D}_2) - q(\mathcal{D})| \leq \chi + \gamma.$$

Let S_1 be the coordinates corresponding to the accepted samples and S_2 be the coordinates corresponding to the rejected ones. The key observation is that $S = S_1 \cup S_2$ is subset generated by sampling from the uniform distribution and has a distribution supported on it that approximates $\mathcal{D}$. So, it suffices to bound the size of S. The probability that a given sample gets accepted is

$$\sum_{x \in \mathcal{X}} \frac{\mathcal{D}_1(x) N\sigma}{N} = \sigma.$$

Thus the expected number of samples needed in the rejection sampling procedure is $M = O\left(\frac{\log(2/\eta)+d}{\sigma\gamma^2}\right)$. Using a Chernoff bound, we can bound the probability that this is greater than its mean by a factor of 4 by

$$e^{-M} \leq \frac{\eta}{2}$$

where we used that fact that $M \geq \log(2/\eta)$. □

We are finally ready to prove our theorem.

Theorem G.4. *For any data set B that is (σ, χ)-smooth with respect to a set of queries $\mathcal{Q}$ of VC dimension d, the output $\mathcal{D}''$ of the Subsampled Net Mechanism satisfies that with probability $1-\eta$, for all $q \in \mathcal{Q}$*

$$|q(\mathcal{D}_B) - q(\mathcal{D}'')| \leq \chi + \tilde{O}\left(\sqrt[3]{\frac{d\log(1/\sigma) + \log(1/\eta)}{\epsilon n}}\right)$$

Proof. Consider a subset V sampled with size $M = O\left(\frac{\log(1/\eta_1)+d}{\sigma\gamma^2}\right)$ where η_1 and γ are parameters we will set later. From Lemma G.3, with probability $1-\eta_1$ we have that there exists a distribution $\mathcal{D}'$ supported on V such that for all $q \in \mathcal{Q}$

$$|q(\mathcal{D}') - q(\mathcal{D}_B)| \leq \chi + \gamma.$$

Let us work conditioned on this event. Let A denote the set of all data sets supported on V and let C denote all data sets supported on V with size $O\left(d\gamma^{-2}\right)$. From Fact G.2, for any distribution $\mathcal{D}_1$ supported on V, there is a data set in C whose distribution $\mathcal{D}_2$ satisfies

$$|q(\mathcal{D}_1) - q(\mathcal{D}_2)| \leq \gamma.$$

We recall the guarantees of the exponential mechanism (see e.g. **?**): Let B'' be the data base output by the exponential mechanism. Then,

$$\Pr\left[s(\mathcal{D}_{B''}, \mathcal{D}_B) \geq \min_{B_1 \in C} s(\mathcal{D}_{B_1}, \mathcal{D}_B) - \frac{2}{\epsilon n}(\log|C| + t)\right] \leq e^{-t},$$

where $s(\mathcal{D}_B, \mathcal{D}_{B'}) = \max_{q\in\mathcal{Q}} |q(\mathcal{D}_B) - q(\mathcal{D}_{B'})|$. Note that $\log|C| \leq M^{O\left(d\gamma^{-2}\right)}$. Thus, with probability $1-\eta_2$,

$$s(\mathcal{D}_{B''}, \mathcal{D}_B) \geq \min_{B_1 \in C} s(\mathcal{D}_{B_1}, \mathcal{D}_B) - \gamma$$

for

$$\gamma \geq \frac{4}{\epsilon n} \log \frac{M^{O(d\gamma^{-2})}}{\eta_2}.$$

Since, $\min_{B_1\in C} s(\mathcal{D}_{B_1}, \mathcal{D}_B) \leq \chi + 2\gamma$, setting $\eta_1 = \eta_2 = \eta/2$ and solving for γ, we get

$$\gamma = \tilde{O}\left(\sqrt[3]{\frac{d\log(1/\sigma) + \log(1/\eta)}{\epsilon n}}\right)$$

as required. □

G.1 Running Time of Subsampled Net Mechanism

The running time of the algorithm involves first sampling M elements uniformly from the domain which takes time $O(M \log|\mathcal{X}|)$. Each query needs to be evaluated on the data set B which takes time $n|\mathcal{Q}|$. Evaluating and sampling from all data bases as required by the exponential mechanism naively takes time $M^{O\left(d\gamma^{-2}\right)}$. As discussed earlier, this can be sped up using sampling for approximation.